中国软件杯
作品简介封面

项目名称：A8——智能简历解析系统

参赛队伍：猫猫狗狗队

所属院校：苏州大学计算机科学与技术学院

队　　长：蔡博凯

队　　员：郭旭东、谢欣欧

指导老师：钱忠

目录

1 项目概述.....3
2 创新点和亮点.....3
2.1 深度学习模型.....3
2.2 科学的评价体系.....4
2.2.1 PRF 评估指标.....4
2.2.2 层次分析法.....4
2.3 完善的系统功能.....6
2.4 多维度特征提取.....8
3 技术架构.....9
4 具体功能实现.....10
4.1 上传文件.....10
4.1.1 上传文件的种类介绍.....10
4.1.2 单文件上传与多文件上传辨析.....10
4.2 解析文件.....10
4.3 信息展示.....10
4.3.1 个人信息展示.....11
4.3.2 统计信息展示.....11
4.4 人岗匹配.....11
5 挑战与克服.....12
5.1 文件上传.....12
5.2 简历解析.....12
5.3 调用模型.....12
6 未来展望.....13
参考文献.....13

图表 1 不同模型对比情况.....4
图表 2 BERT_BiLSTM_CRF 模型在简历数据验证集上的 PRF 值.....4
图表 3 递阶层次结构模型.....5
图表 4 岗位匹配权重分布雷达图.....6
图表 5 上传文件功能简介与技术.....6
图表 6 简历解析功能简历及技术.....7
图表 7 岗位匹配各特征匹配度计算.....7
图表 8 信息展示模块与跳转逻辑.....7
图表 9 提取字段介绍.....8
图表 10 技术架构.....9

1 项目概述

我们的项目是一款基于人工智能技术的简历解析系统，旨在帮助企业高效地筛选和分析大量的求职者简历。当前的招聘流程中，处理大量简历的问题一直是一项挑战。我们的简历解析系统采用了先进的自然语言处理和机器学习技术，可以自动从文本简历中提取关键信息并转化为结构化的数据，为招聘人员提供更快速、准确的招聘决策。

2 创新点和亮点

2.1 深度学习模型

本系统采用了百度的开源词法分析工具 LAC[1]，主要原理在于通过分词将文本分块，然后再进行各词语的命名实体识别(NER)，通过这个工具可以提取文本中的一般信息，比如姓名。

本系统采用了百度 PaddleNLP 的 Taskflow 调用通用信息抽取(UIE)模型[2]，由于该模型已经在大样本下进行了初步训练，因此使用该模型进行一般信息的抽取准确率也很高，同时可抽取的信息也很广泛，比如姓名、毕业院校、学历、籍贯、职业、公司等信息。

通过对于已抽取信息的观察，发现 UIE 模型在一些与简历强相关的信息比如职业、学历、公司上的提取精度上来说还有优化的空间。与此同时，可以发现 LAC 工具处理句子是采用的是先分词，然后在对词语进行标注的方法，由于少数情况下一个句子可能会产生多种分词方式，这样会对 NER 任务产生影响。

因此可以采用一种新的模型，以单个字符作为单位进行标注，同时采用简历命名实体识别数据集[3]对模型进行微调。这个模型就是 BERT_BiLSTM_CRF 模型[4]，BERT 模型保证将一个字符映射到唯一的 id，通过对输入的 id 进行训练最后可以达到理想的 NER 识别效果。这个模型在姓名、学历、职业、公司等信息的识别上都有很好的效果。

以下是三种模型在姓名和学历信息上的提取正确率，数据集使用比赛提供的

100 份简历。可以看到 BERT_BiLSTM_CRF 模型和 UIE 模型略优于 LAC 模型，因此在系统中采用前两种模型参与信息抽取。

	LAC	UIE	BERT_BiLSTM_CRF
姓名	97%	98%	97%
学历		97%	99%

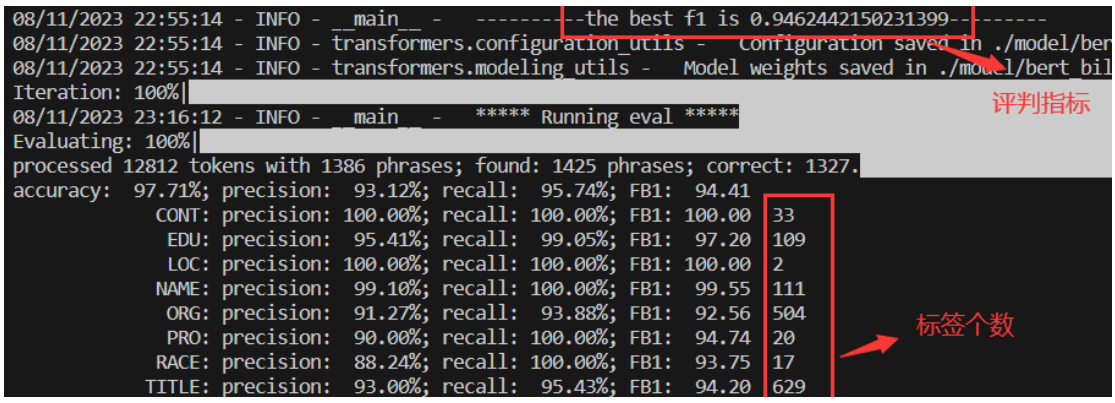
图表 1 不同模型对比情况

2.2 科学的评价体系

2.2.1 PRF 评估指标

在评估 BERT_BiLSTM_CRF 模型时采用 PRF 评估指标，PRF 值分别表示准确率（Precision）、召回率（Recall）和 F1 值（F1-score）。准确率是指在预测正确集合中真正正确的比例，召回率是指在真正正确的集合中预测正确的比例，F1 值可通过 PR 值计算，这是二分类问题的基本原理，在该模型中需要解决多分类问题，可最直接计算各类的 PR 值，最后通过加权平均计算 F1 值。

下图是模型中各类标签在验证集上的 PRF 值。

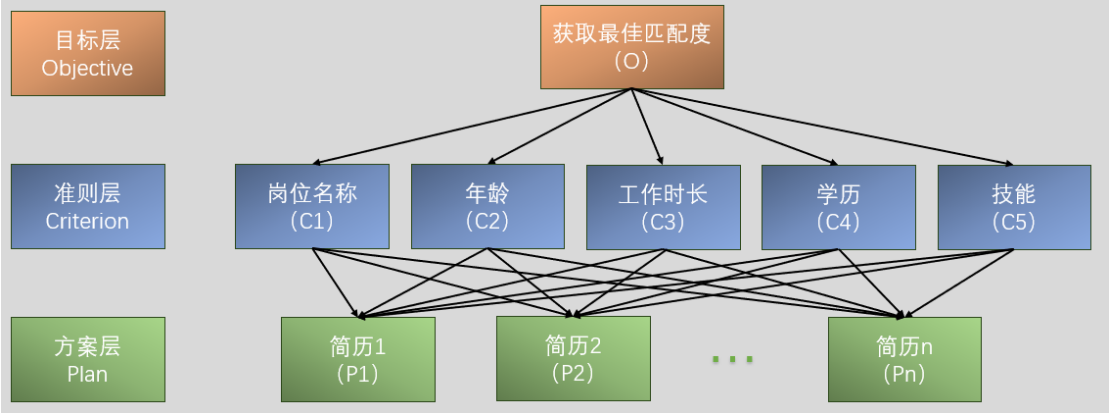


图表 2 BERT_BiLSTM_CRF 模型在简历数据验证集上的 PRF 值

2.2.2 层次分析法

在进行岗位匹配的任务时，需要确认不同特征对匹配任务的重要性，以此计算各个特征的权重，这个过程可以通过层次分析法来完成，以下是具体步骤。

① 建立递阶层次结构模型



图表 3 递阶层次结构模型

② 构造判断矩阵

表格 1 判断矩阵

	岗位名称	年龄	工作时长	学历	技能
岗位名称	1	5	3	2	7
年龄	1/5	1	1/2	1/3	2
工作时长	1/3	2	1	1/2	3
学历	1/2	3	2	1	2
技能	1/7	1/2	1/3	1/2	1

③ 一致性检验

一致性检验：

最大特征值λ_{max}=5.1237

指标数n=5

一致性比例CR=$\frac{CI}{RI}$

CI=$\frac{\lambda_{max}-n}{n-1}=\frac{5.1237-5}{5-1}=0.03$

RI通过查表可知为1.12

CR=0.0276 < 0.1 通过一致性检验

④ 计算权重

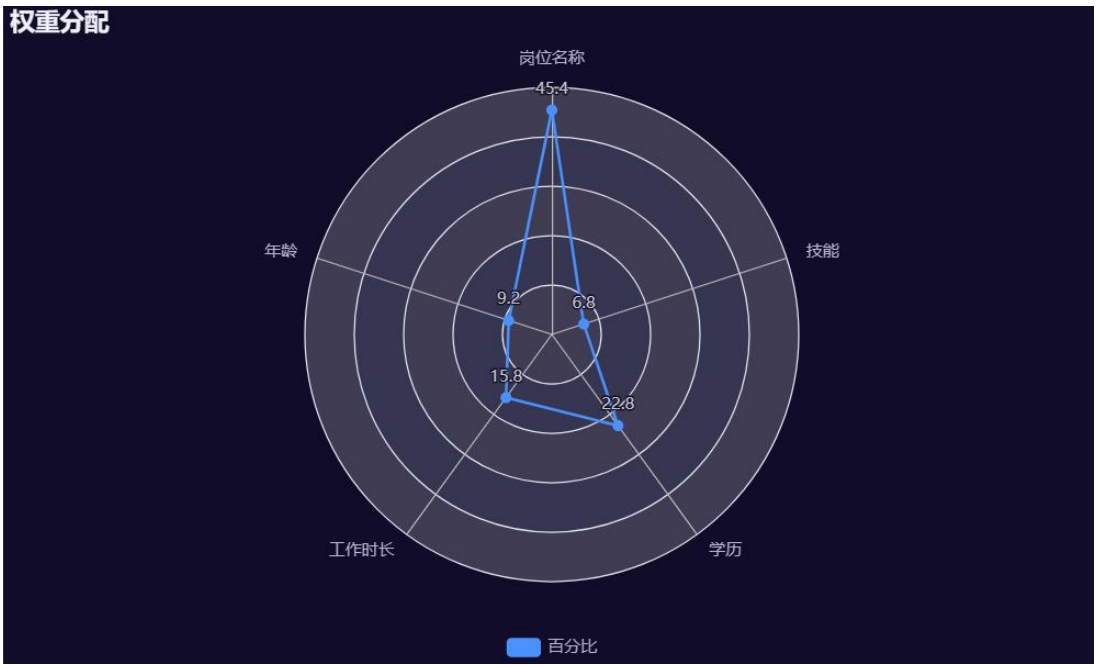
计算权重（算术平均法）：

a. 将判断矩阵按列归一化

b. 将归一化的各列相加

c. 将相加后的向量数乘1/n

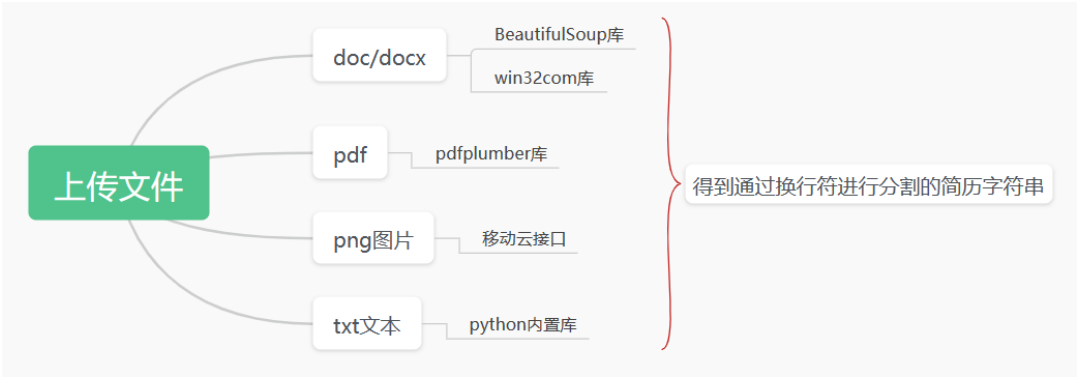
最后将得到的权重绘制成雷达图。



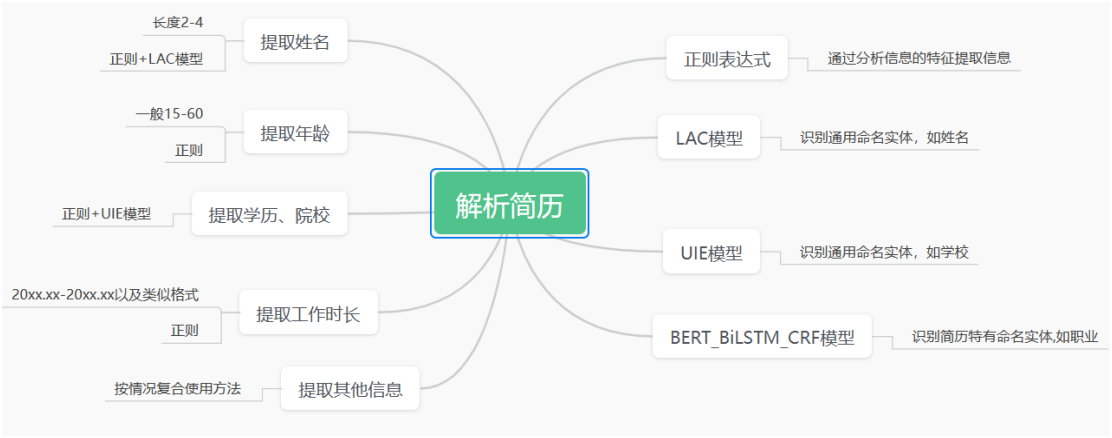
图表 4 岗位匹配权重分布雷达图

2.3 完善的系统功能

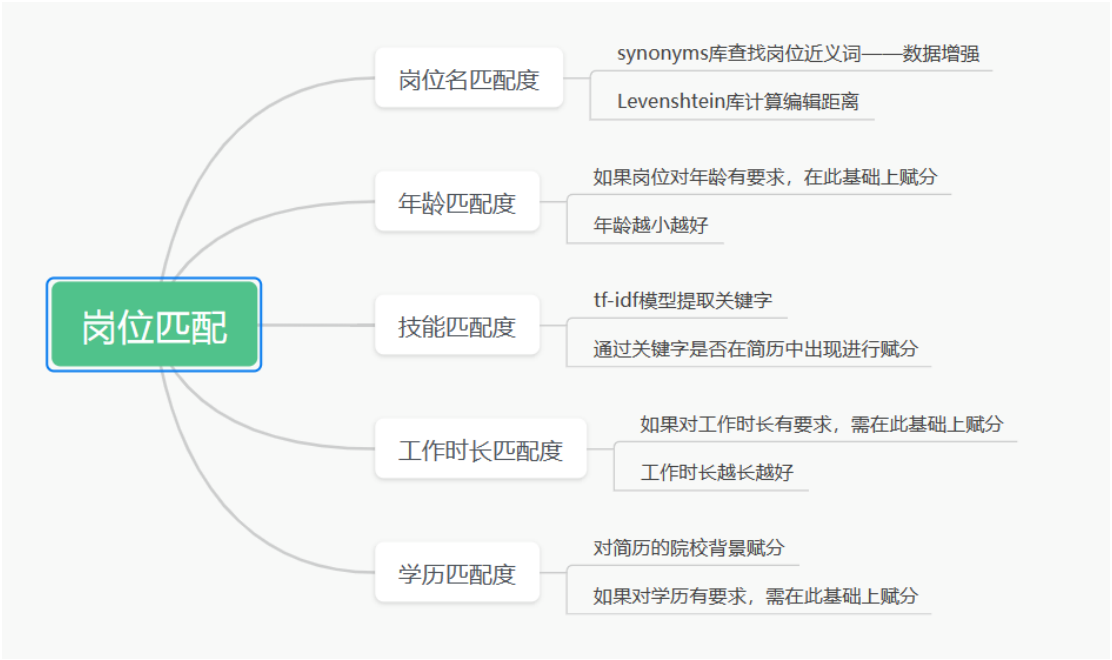
本系统的系统功能完善，功能模块主要包含文件上传、解析文件、人岗匹配、信息展示。下面介绍不同功能所使用的技术。



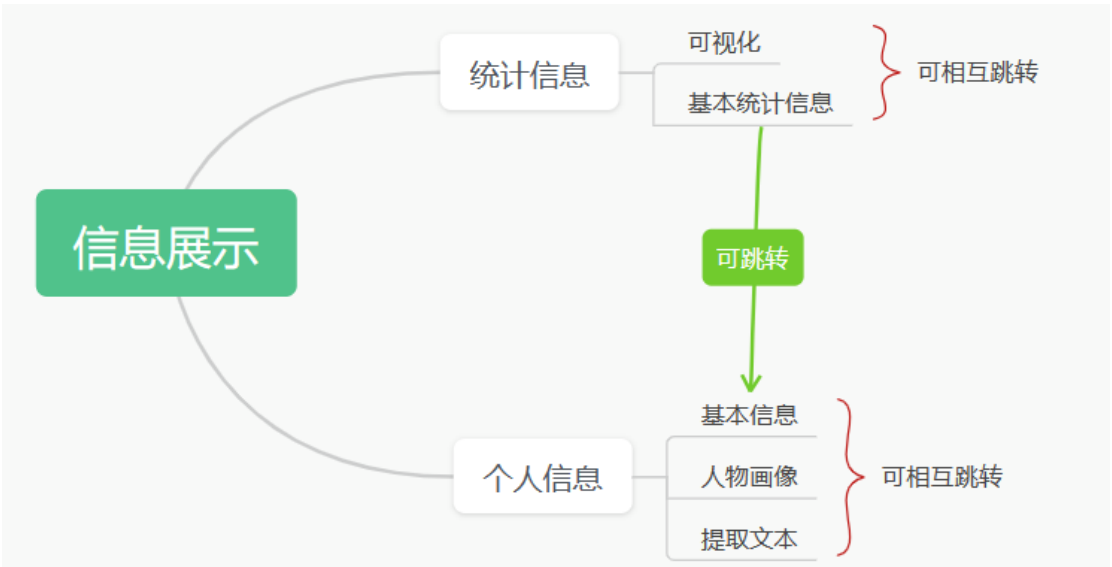
图表 5 上传文件功能简介与技术



图表 6 简历解析功能简历及技术



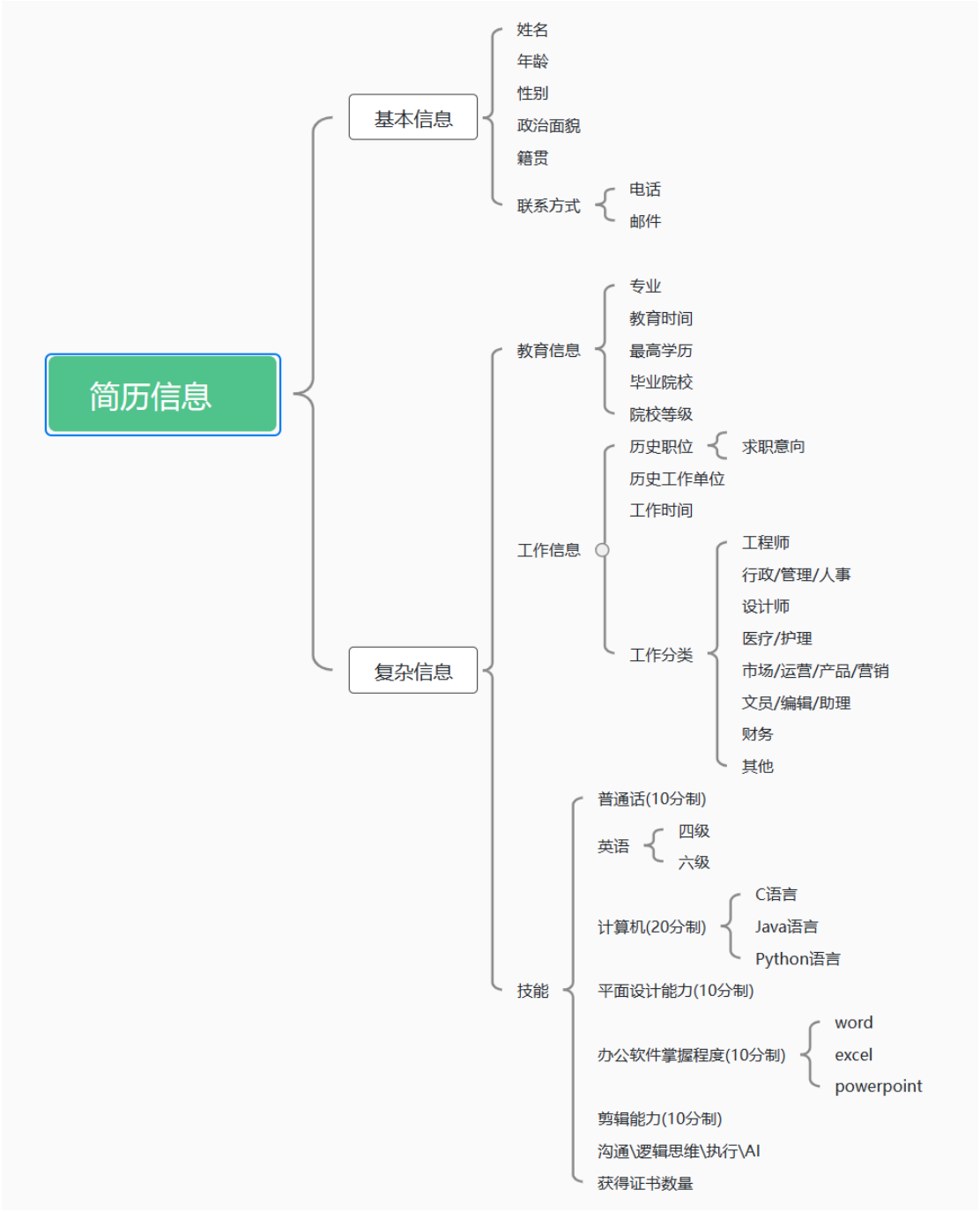
图表 7 岗位匹配各特征匹配度计算



图表 8 信息展示模块与跳转逻辑

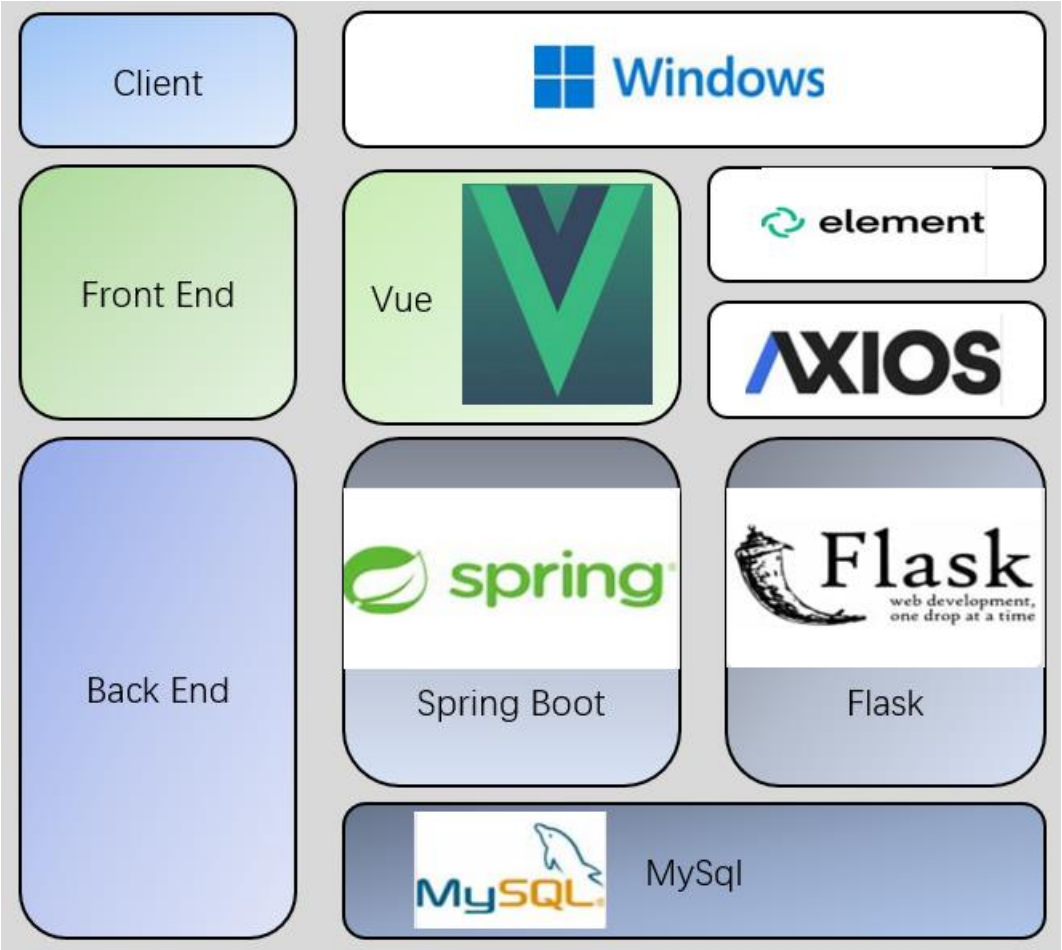
2.4 多维度特征提取

在解析简历这个功能模块中，不仅包含比赛要求的五个基本信息(姓名、年龄、学历、院校、工作时长)。而且还扩展了一部分信息，最后可利用的字段达到 40 余个。如下图所示。



图表 9 提取字段介绍

3 技术架构



图表 10 技术架构

4 具体功能实现

4.1 上传文件

4.1.1 上传文件的种类介绍

该功能支持上传 txt 文档，doc 文档，docx 文档，图片和 pdf，其中可以直接读取 txt 文档中的内容，如果是 doc 文档就把它转化为 docx 文档进行处理，如果是 docx 文档就使用 win32com 包提取内容，如果是 pdf 就使用 pdfplumber 包提取内容，如果是图片的话就使用移动云提供的包提取内容。并且将这些提取出的字符串按指定格式使用换行符分割。

4.1.2 单文件上传与多文件上传辨析

单文件上传与多文件上传是在逻辑上将上传功能分为两个部分，单文件上传表示上传 n 份文件，就把这 n 份文件看做是一份简历进行处理；多文件上传则表示上传 n 份文件，就把这 n 份文件看做是 n 份简历进行处理

4.2 解析文件

此时系统内部会对上传文件得到的字符串进行进一步处理。使用的主要工具是正则表达式,自己微调的模型 BERT_BiLSTM_CRF,以及 paddlenlp 的 UIE 模型。进行的任务主要是命名实体识别(NER)，比如姓名和籍贯等，或者根据信息的一些基本体征进行正则表达式提取，如电话号码和邮箱等。但是更多的，需要将这些方法结合起来，以提高提取的正确率。

4.3 信息展示

信息展示分为统计信息展示和个人信息展示。

4.3.1 个人信息展示

个人信息展示包括基本信息展示和人物画像，其中基本信息就是将简历解析提取到的信息做一个前端展示。而人物画像则需要通过提取到的信息，对于简历做出评价，这里则需要引入评分机制，根据分数来判断候选者在某字段上的能力。这里使用雷达图展示候选者可能擅长的工作，使用条形图展示候选者的技能信息，通过文字描述对用户进行人物画像的刻画，比如如果用户的“英语能力”为六级的话，会评价该人的英语水平良好，如果用户毕业于 985 或 211 大学的话会评价该人的院校背景不错，这些内容会在人物画像的简历亮点模块予以展示。

4.3.2 统计信息展示

统计信息是包含统计信息展示和信息的可视化展示，其中统计信息展示是指将数据库中的简历按照基本信息一条一条地列出来，同时提供一个详情按钮使得用户可以跳转到对应的个人信息界面，信息的可视化指将数据库中的简历按照特定信息统计各类的人数，同时使用饼状图将这些信息展示出来，比如按照学历将数据库中的简历信息分为 985、211 和其他。

4.4 人岗匹配

人岗匹配是指 HR 通过输入岗位名称、年龄要求、学历要求、工作时间要求、以及详细需求，与数据库中的简历进行匹配。这里介绍一下大致的匹配过程，首先对详细需求进行关键字提取，用 jieba 包中的 extract_tags()函数，进行分词、提取使用默认的 TF-IDF 模型对文档进行分析，同时去除停用词。然后将岗位名称，年龄学历工作时间要求以及关键词列表按照指定权重对简历进行匹配度分析，找到匹配度最高的 20 份简历并将其输出。

5 挑战与克服

5.1 文件上传

在处理文件上传任务时，前端传过来的文件是一个对象，但是一般在处理文件的过程中是按照路径进行处理的，因此，一个难点在于如何将不同种类的对象转化为路径存储到系统中。解决方法在于使用 uploads 文件夹存储文件，同时将不同文件对象根据种类的不同使用各种方法将其写入临时文件夹 uploads 中。在完成后解析简历后会自动删除上传的文件，以防止扰乱之后上传的文件。

5.2 简历解析

在处理文件经过文件上传得到的字符串时，提取到的标签的准确率是一个核心的问题。一开始我们使用正则表达式进行提取，除了年龄可以通过正则表达式达到 100%正确率以外，其他四个基本特征均有不足之处，同时考虑到工作时长的数据特征比较明显，可以通过正则表达式进行进一步提取，另外的姓名、学历、毕业院校均可以通过深度学习模型进行补充提取，最后各特征的正确率如下。其中，工作时长由于一些标注的信息与实际中通过简历人为计算的信息不符合（例如 66.docx/54.docx/51.docx/49.docx/41.docx/32.docx，这些在标注文件中错误地将实习 or 兼职经历算入工作经验中），因此正确率偏低。

表格 2 系统在比赛提供训练集上的正确率

	姓名	年龄	学历	毕业院校	工作时长
正确率	99%	100%	99%	96%	88%+6%=94%

5.3 调用模型

在调用模型的过程中，出现无法找到模型的路径这个问题。原因是采用了相对路径，由于工作路径与正在运行的 run.py 是一个路径，因此在子文件夹中的程序访问相对路径时会造成路径的错误，最后将路径全部改成绝对路径，问题得以解决。

6 未来展望

做完这个系统后，经过测试，发现两个问题，一个是简历解析的时间过长，保守估计在 10s 左右，二是由于设备的限制，难以进行多轮次的训练，以及在并发访问上传文件时只能进行异步上传文件。

之后会在两个方面进行优化，一个是训练出一个统一的模型专门用来提取简历中的信息，而是租服务器对模型进行多轮次的训练，同时实现同步上传文件的功能扩展。

参考文献

[1] https://github.com/baidu/lac

[2] https://paddlenlp.readthedocs.io/zh/latest/model_zoo/taskflow.html

[3] https://tianchi.aliyun.com/dataset/144345

[4] Dai Z, Wang X, Ni P, et al. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records[C]//2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei). IEEE, 2019: 1-5.